View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by International Journal of Digital Curation

IJDC | Research Paper

The Red Queen in the Repository: Metadata Quality in an Ever-Changing Environment

Joakim Philipson
Stockholm University Library

Abstract

One of the grand curation challenges is to secure metadata quality in the ever-changing environment of metadata standards and file formats. As the Red Queen tells Alice in *Through the Looking-Glass*: "Now, here, you see, it takes all the running you can do, to keep in the same place." That is, there is some "running" needed to keep metadata records in a research data repository fit for long-term use and put in place. One of the main tools of adaptation and keeping pace with the evolution of new standards, formats – and versions of standards in this ever-changing environment are validation schemas. Validation schemas are mainly seen as methods of checking data quality and fitness for use, but are also important for long-term preservation. We might like to think that our present (meta)data standards and formats are made for eternity, but in reality we know that standards evolve, formats change (some even become obsolete with time), and so do our needs for storage, searching and future dissemination for re-use. Eventually, we come to a point where transformation of our archival records and migration to other formats will be necessary. This could also mean that even if the AIPs, the Archival Information Packages stay the same in storage, the DIPs, the Dissemination Information Packages that we want to extract from the archive are subject to change of format. Further, in order for archival information packages to be self-sustainable, as required in the OAIS model, it is important to take interdependencies between individual files in the information packages into account. This should be done already by the time of ingest and validation of the SIPs, the Submission Information Packages, and along the line at different points of necessary transformation/migration (from SIP to AIP, from AIP to DIP etc.), in order to counter obsolescence.

This paper investigates possible validation errors and missing elements in metadata records from three general purpose, multidisciplinary research data repositories – Figshare, Harvard's Dataverse and Zenodo, and explores the potential effects of these errors on future transformation to AIPs and migration to other formats within a digital archive.

Received 14 January 2019 ~ *Accepted* 13 August 2019

Correspondence should be addressed to Joakim Philipson, Stockholm University, SE-10691 Stockholm, Sweden. Email: joakim.philipson@su.se

An earlier version of this paper was presented at the 13th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Copyright rests with the authors. This work is released under a Creative Commons Attribution Licence, version 4.0. For details please see https://creativecommons.org/licenses/by/4.0/

Introduction

To meet high quality metadata standards of ingested documents, a research data repository must constantly adapt, evolve, and refine upload methods and export formats to meet the demands of the depositors and other stakeholders, such as the potential re-users of data, present and future, thereby also contributing to answering the call from the reproducibility crisis in science (Kingsley, 2018). Research (meta)data, in order to be FAIR, should not only be findable and accessible, but also interoperable and re-usable. For this purpose, it is insufficient merely to carefully reproduce files and metadata records. A simple copying of original input metadata is not enough. To continue being re-usable and fertile also in the future, reproduction in a data repository needs to involve also a recombination of metadata elements and enrichment with e.g. preservation metadata (such as PREMIS or PROV), documenting the origin (provenance) of (meta)data and the possible changes (meta)data have gone through in the repository. This is necessary simply in order to survive and be relevant for generations to come.

The responsibility for making metadata comply with the FAIR-principles lies to the largest part with the repositories. Assigning *Persistent Identifiers* (PIDs) such as DOIs, to datasets and displaying them well in landing pages or item records, for easy citation, is evidently a task incumbent on repositories, serving to make datasets (items) *findable* and *accessible*. Managing metadata standards, including export formats, and licensing, e.g. by means of drop-down menus to choose from, are also responsibilities of repositories, contributing to make datasets (items) more *findable*, *interoperable* and *re-usable*. Most important, repositories are the key players in making metadata *machine actionable* or at least machine readable, "a *conditio sine qua non* for FAIRness."[1] A number of digital resources, some of them in repositories which are part of this study (Harvard's Dataverse, Figshare, Zenodo), were evaluated according to the *fairmetrics.org* measures,[2] and none of them scored a full 100% on the 16 measures. (Dataverse seems to have "faired" best among them, though, with four "yellow cards", meaning "problematic", one "reddish", signalling here a failed FAIR metric due to "no response provided", and one "grey", meaning "cannot be evaluated"). This result was obtained despite the obvious temptation for participating resources or repositories to pick their "best shot" for evaluation, as the choice was entirely theirs to make. Adherence to the FAIR principles of *interoperability* and *re-usability* could be facilitated by repositories providing upload web forms with inherent pre-ingest validation conformant to some general metadata standard, including datatypes. Tooltips in such web-forms could make it easier for the individual uploading researcher to do the right thing. It is further the responsibility of repositories to see to that their metadata output conform to all the export metadata formats and standards that they profess to provide. This has not always been so. There have been interactions directly with repository staff in at least two cases, to make them remove systematic validation errors and improve export format output to conform to said metadata standard. In one case, it concerned basic things such as making the METS xml-file becoming well-formed, by removing prefixes from attributes and including the missing, mandatory element *structMap* in the file. In the other case, it concerned the failure to validate against DDI codebook 2.5,[3] still an unresolved issue, as seen from this study. For validation to be part of the solution to the challenge posed by

1 See: https://www.force11.org/fairprinciples#Annex6-9
2 fairmetrics.org measures: https://doi.org/10.5281/zenodo.1305060
3 See: https://github.com/IQSS/dataverse/issues/3648

constant evolution of formats and standards it must be kept in mind that validation schemas change along with the formats that they define. This means that validation schemas and transformation code, should also be archived together with the AIPs. Li and Sugimoto have argued extensively that:

> "Metadata schema changes may cause inconsistency in the use of metadata, which is also a risk for the long-term use of digital resources. Due to the high cost of re-creation of metadata, longevity of metadata is an important issue for long-term use of digital resources. Metadata schema, which defines a set of terms, structure of metadata instances and some related characteristics of metadata instances, has to be maintained as well as the metadata instances over time" (Li and Sugimoto, 2014).

and that:

> "Long-term maintenance of metadata schemas and metadata vocabularies is [a] significant issue for keeping metadata interpretable over time" (Li and Sugimoto, 2018).

By ensuring compliance with standards, these tools are essential in controlling uniformity of records in a heterogeneous collection, for future needs of transformation and migration to new, sustainable formats. The original object of this paper was to give examples of validation errors of metadata encountered in four general purpose, multidisciplinary repositories or platforms for the publication of research datasets, Eudat's B2Share, Figshare, Harvard's Dataverse and Zenodo. However, it soon turned out that Eudat's B2Share was less suitable for this purpose, since the output formats offered did not allow meaningful comparison with the other repositories, also since the metadata records within that repository were not sufficiently uniform, often having their own individual formats and validation schemas.

For these reasons, it was decided to leave Eudat's B2Share out of this study. By contrast, Dataverse, Figshare and Zenodo are offering a variety of export metadata formats, but do not always comply with the standards of these proposed formats. These kinds of errors, sometimes occurring due to a lack of effective input constraints and validation at ingest, might cause problems in transforming even relatively small collections of items with substantial variations between them.

Even when compliance with metadata standards does occur, for some export formats there is still a lack of file metadata, such as formats, sizes, checksums, mime-types and even original names of the files comprising the dataset. Such file metadata may be important for the further identification, processing and transformation of SIPs (submission information packages) into Archival Information Packages (AIPs) in compliance with the OAIS model – the Open Archival Information System reference model (CCSDS, 2012). AIPs often need enrichment with preservation metadata, such as PREMIS events, for which such file metadata are important parameters.

The AIP metadata files are largely generated through transformation and enrichment of the corresponding SIP metadata files, what in the OAIS reference model is called the Packaging Information, of which there is "exactly one piece" that "identifies and delimits the Information Package."[4]

Further, in the OAIS model is described a *Develop Packaging Designs and Migration Plans* function which "receives Archive approved standards and migration goals from

4 ibid., p. 4-33

Administration," including "format standards, metadata standards and documentation standards. It applies these standards to preservation requirements and provides AIP and SIP template designs to Administration. This function also provides customization advice and AIP/SIP review to Administration on the application of those designs."[5] This is also what metadata validation schemas do, defining these standards and evaluating conformance of information packages received to them.

The OAIS model is perhaps most clearly embodied in the METS standard,[6] with its different sections of *dmdSec*, *amdSec* (with subsection techMD), *fileSec, structMap* for descriptive ("bibliographic"), administrative and technical, file metadata and structural relationships between data files. All of these sections may contain expressions of different metadata standards and be subject to validation against several different metadata schemas. (METS is a metadata "wrapper" standard, which allows using different suitable metadata formats or standards in different sections. For e-legal deposit at the National Library of Sweden, for example, MODS is used in the *dmdSec*, while PREMIS is used for preservation relevant information in the amdSec/techMD section of the AIPs).

Long-Term FAIR-ness of Research Datasets?

At the International Digital Curation Conference 2018 in Barcelona, a talk on long-term preservation of research data posed a highly relevant question: are research data sets FAIR in the long run? (Wehrle and Rechert, 2018). The FAIR principles, (Findable, Accessible, Interoperable and Re-usable) do not as yet contain any formal technical requirements, while long-term preservation of research data are heavily dependent on technical format of files, software, hardware. The investigation underlying the talk was based on a selection from *Re3Data* of 92 repositories, comprising 3.5 million data files (1.95TB), in order to find out to what extent they were sustainable and fit for long term preservation. For the analysis they used Harvard's FITS, an open source software container for a number of tools aimed at file format identification or validation (e.g. DROID, JHOVE etc.), and *eCommons* from Cornell's Digital Repository,[7] which has a "probability" table for the long-term sustainability of different file formats for various media types (audio, video, images, text, spreadsheet.).

Lessons Learned

Some important general lessons learned from that study are that research datasets often are quite heterogeneous, with strong interdependencies between file formats in a fileset of several datasets. (A fileset might be conceived as a micro ecosystem in itself, where changes in one file might affect relationships between files throughout the whole fileset.) This means one cannot simply focus on preservation and migration of individual files to more sustainable file formats, but must take entire filesets or information packages in to account.

The FAIR-principles proved to be important for long-term preservation; despite being abstract, they nevertheless tend to foster long-term hindsight. Handling of different research data file formats for long-term preservation (e.g. by validation,

5 ibid., p. 4-15
6 METS Standard: http://www.loc.gov/standards/mets/mets-home.html
7 eCommons Recommended File Formats: http://guides.library.cornell.edu/ecommons/formats

conversion/transformation) may never be fully automatable, but always require some manual efforts.

Object and method

The question arises from the Wehrle and Rechert study, whether the same lessons learned could be applied also to metadata files or records, which after all – one might think – should be less complex and more homogeneous. Nevertheless, at least some metadata schemas are quite comprehensive and complex, as e.g. the Data Documentation Initiative (DDI)[8] used by Harvard's Dataverse and some other repositories (that are not part of this study).

The Wehrle and Rechert study is based on, as a first step, a survey of difficulties in analyzing technical characteristics (file formats) in real life research data. For metadata, it is rarely the file format of the metadata files as such (often XML or JSON) that is the problem. The descriptions or definitions of the metadata standards themselves, however, are given by specifications and validation schemas, which are thus instrumental tools for metadata analysis. Some repositories also allow for customized metadata (which may not be part of export format outcomes, though) or a "mix and match" of metadata standards, that make records more heterogeneous, but these were not included here.

In this study, to investigate the possible future effects of erroneous metadata on future archival processing a small sample was made of metadata records from three multi-disciplinary, i.e. not domain-specific repositories (dataverse.harvard.edu, figshare.com, and zenodo.org).

From these, a selection both from the most currently published metadata records (from 2018), and from those published 2015 or before, were validated against their inherent schemas (as given in the *schemaLocation*), if present. In case there was no *schemaLocation*, the "nearest suitable" validation schema as indicated by the namespaces in the record was used. For Dataverse *dcterms* records, it was actually necessary to construct a new complementary validation schema, *metadataDCT.xsd,* that imports the regular *dcterms.xsd* schema while also defining the container, root element <metadata/>, otherwise lacking in the general Dublin Core metadata standard. The two groups were checked for metadata standard version changes in the time lapse between "older" and current records.

The object of the study was to detect possible validation errors or missing data and find out how these might affect possible transformations to archival format (AIP) for long-term preservation. To perform validation, Oxygen XML Editor 19.0 with schema validation engine *Xerxes* and default XML schema version 1.0 was used.

Expected Outcomes

The resulting dataset from this study contains lists of URLs or URIs for the OAI-PMH feeds, alternatively for individual, single items or metadata records, when export format metadata records otherwise could not be harvested and validated in XML. There are also references to metadata standards, their different versions and associated validation schemas, with types of validation errors encountered in the samples and their potential effect on future transformation and migration efforts in some cases. This, naturally, will sometimes be difficult to project, since we cannot anticipate fully what standards and formats will be current in the future.

A general preliminary hypothesis is that the more homogeneous a collection of metadata records is, the more easily it will be subject to transformation and migration.

8 DDI: http://www.ddialliance.org/

Paradoxically, systematic errors affecting all metadata records in a collection in the same way may be more easy to handle, e.g. by means of using other metadata sources for the affected elements.

However, the level of homogeneousness of metadata output records will also depend on the restrictiveness of the validation schemas and in particular, the validation performed pre-ingest. The more lax a validation schema, with little or no value content or datatype control (as for *dcterms.xsd*), the more room is open for a heterogeneous metadata output, which might cause trouble later on in the process of data processing and transformation.

We experienced the possible impact of heterogeneous metadata, with no validation of content values at ingest, when in 2014 at the National Library of Sweden, there was a small pilot conversion project of transforming a set of personal archival records from EADXML[9], (the Encoded Archival Description, maintained by the Library of Congress), to the bibliographic standard of the Swedish National Union Catalogue, then using a variant of the MARC21 format, LIBRISMARC.

On the face of it, this looked as a rather straightforward process and an easy task, transforming one well-regulated XML-format to another. Although the final result proved to be acceptable, with a total of 724 new records of manuscript collections in the union catalog,[10] the process of getting there took longer time and a more strained effort than what could be imagined at the outset.

Problems involved for example the variations in formatting of personal names in EDIFFAH (a common Swedish database for personal archives of manuscripts, letters etc.); sometimes family names were comma separated from first names, sometimes not, sometimes with years of lifespan within parentheses, sometimes not etc.

Another problem was the varying formatting of time intervals covered by a collection. The complexity and variation in formatting content of values only of "cover years" in the original EAD database was reflected in the rather unwieldy XSLT stylesheet used for the final transformation then, and the multiple variables that were needed to handle only this one element in the XML-file. A greater uniformity of input (ingested) data would have made it much simpler. This could have been achieved by a more restrictive input validation of datatypes and content values, already at pre-ingest.

Thus, one way of meeting the challenge of heterogeneous metadata, the evolution of metadata standards and the necessity of transformation and future migration to new, yet unknown metadata standards, might be to impose more strict validation schemas, and perhaps even more important, to perform mandatory pre-ingest validation. Web forms for upload of data should offer tooltips and guidance helping data providers to do the right thing and format their metadata to comply with standards and schemas.

Another possible solution to the challenge posed by evolution of metadata standards and formats, could be a closer collaboration or integration between possible metadata sources (such as applications for funding, data management plans, ethical vetting documents etc.), involving agreement on common identifiers (PIDs) e.g. for projects, persons and organisations, and making these potential metadata sources truly machine actionable. This might allow research data managers, repository keepers and archivists to run towards a common, albeit partly hidden goal, to keep pace with the evolution of new metadata standards and file formats.

9 EADXML: https://www.loc.gov/ead/

10 See: http://libris.kb.se/hitlist?q=db%3aHARK&d=libris&m=10&p=1&s=r

Selection of datasets (items) for validation of metadata files

One objective of the test validation of item metadata files was to capture possible changes occurring over time, as a result of version changes in metadata standards in the course of repository development. For this reason, two sets of items from each repository were selected for the test collection, one set with publication date from 2018, sorted by the latest publications first, presumably holding the most accurate metadata files. The second set selected for the test collection holds items published in the repositories already in 2014 and 2015, aimed for comparison with the first set and tracking possible changes occurring in the period between the two sets.

Metadata standards for validation and repositories

The metadata standard compliance that is tested differs between the repositories in this study, which makes the comparison somewhat difficult. This is partly because the complexity and restraints of different metadata standards vary substantially. For the scope and timeframe of this study, limits were also set by the methods of retrieval of test sets and the metadata export formats offered as XML by different repositories, and the presence or absence of general metadata validation schemas for these formats.

One result of these constraints was that, contrary to the intent of the original abstract, soon after discovering that EUDAT metadata records were only available as JSON and with individual tailor-made validation schemas for each item, making comparison with metadata records from other repositories less feasible, we decided to leave EUDAT out of the selection made for this paper.

This left us with a selection of metadata records from three general purpose research data repositories to be tested and validated, Harvard's Dataverse[11], Figshare[12], and Zenodo[13], as indicated above. Records of two or three metadata standards from each repository were tested and validated, for comparison at least one standard being an expression of Dublin Core, the other perceived as representing more or less the "preferred" or default metadata standard of that repository, potentially offering its highest quality records. For Dataverse it was DDI codebook, for Figshare and Zenodo it was Datacite.

Validation Error Types

To ensure a just comparison of outcomes from different repositories, involving different methods for collection of metadata records, only validation errors pertaining to the main descriptive metadata standards count. Hence, validation of metadata retrieved by means of APIs for OAI-PMH feeds in this study disregards minor compliance failures to the OAI-PMH standard, which may sometimes be due to the search-API itself, as in the case of Figshare and Zenodo.

The error type classification used here, with some instances as examples, are described below in Table 1.

11 Dataverse: http://dataverse.harvard.edu
12 Figshare: http://figshare.com
13 Zenodo: http://zenodo.org

Table 1. Validation error types and instance examples.

errorType Codes	errorType Explanation	errorType Instance	objectID Instance
V	**v**oid element (missing value)		
A	**a**ttribute misplaced	Attribute 'URI' is not allowed to appear in element 'producer'.	dvn2018ddi-DO3MSH
M	**m**issing element	[**M**issing parent element **<rightsList>**<rights>CC BY 4.0</rights>**</rightsList>**]	fig2014datacite
C	**c**ontent invalid	Element 'useStmt' cannot have character [children], because the type's content type is element-only.	dvn2018ddi-0KIRBJ
O	**o**rder of elements (misplaced)	Invalid content was found starting with element <relPubl>. One of <sumDscr> …	dvn2014ddi-27595
D	**d**atatype value	The value 'DVN' of attribute 'source' on element 'verStmt' is not valid with respect to its type, '#AnonType_sourceGLOBALS'.	dvn2018dct-4ICF6W
E	**e**numeration value	<verStmt source="DVN">: Value 'DVN' is not facet-valid with respect to enumeration '[archive, producer]'.	dvn2014ddi-27595
S	attribute mi**s**sing		

Two of the error types in Table 1, **V** and **S**, showed no instances in the samples for this study. Both of these are more likely to appear when using validation schemas with stricter content control, so their absence here may be an indication of the relative laxity of the metadata standards under scrutiny in this study.

Some errors identified by the validation engines follow inevitably from others, e.g. a datatype error (D) following from a failure to match values in a given enumeration list (E). Such errors are not counted separately, since they may be corrected by the same action as the first error. When determining the number total of errors in a record or a feed, what counts is the number of separate corrections to perform in order to make it validate.

For example, the following two errors, almost universal in the Zenodo *oai_datacite* feeds, although strongly interrelated, will both count as separate, since it takes two separate correction acts to remove them:

(i) Value 'https://zenodo.org/record/1666965' is not facet-valid with respect to pattern '10\..+/.+' for type 'doiType'.

(ii) Value 'URL' of attribute 'identifierType' of element 'identifier' is not valid with respect to the corresponding attribute use. Attribute 'identifierType' has a fixed value of 'DOI'.

Here, (i) describes an error regarding the **element** *content* (**C**) value, which must comply with a certain pattern for DOIs, while (ii) can be seen as a failure of the **attribute** value to match a certain *enumeration* (**E**) list (consisting of only one value, in this case), alternatively as a *datatype* (**D**) error.

Impact of metadata validation failure on transformation

For the error types in this study, the effects on potentially needed transformation to another metadata standard or format in the future were briefly evaluated.

The **V**- and **S**- errorTypes, of which no instances were found in this study as indicated above, may naturally have the effect of propagating missing values also to the transformed target metadata file.

As for the **A**-type of error, provided the misplaced attribute value is not required elsewhere, it will probably have little or no impact on the potential transformation result.

The **C-, D-** and **E**–types of error might be more severe in terms of creating non-uniform, heterogeneous value content, possibly even causing parsing failures for transformation files (xslt).

The **O**-errorType may cause XPath expressions in transformation files to fail in finding their correct target and causing missing values in the transformation outcome, or possibly also failed parsing. But the risk of that happening may be lower if the misplacement of an element is still in a sequence under the same parent element as prescribed by the schema (given that there are no crucial interdependencies between "siblings" under that same parent, that is).

The **M**-errorType, finally, may potentially also cause some of these adverse effects in transformation efforts, certainly if the missing element happens to be mandatory according to the requirements of the schema(s) involved. But, as noted about the Figshare instances with a missing parent element, given (paradoxically!) that these errors seem to be systematic, it is conceivable that a transformation file could relatively easily be designed, with preserved uniformity of data, with an XPath-expression that simply bypassed the missing parent element and targeted the child directly.

Results

Some of the results of the validations undertaken are found in the figures below. The full datasets, including links to metadata records and validation schemas used, is available in Zenodo.[14]

14 doi:10.5281/zenodo.2276777

Dataverse

objectID	errorTypes	errorCount	note	repository
Dataverse				
dvn2014dct-26792	none	0	none	Dataverse
dvn2014dct-27595	none	0		Dataverse
dvn2014dct-28530	none	0	none	Dataverse
dvn2014ddi-26792	A(1), C(1), O(3), E(1)	6	plus D (1) sameAs E(1)	Dataverse
dvn2014ddi-27595	A(1), C(1), O(3), E(1)	6	plus D (1) sameAs E(1)	Dataverse
dvn2014ddi-28530	0(3), E(1)	4	plus D (1) sameAs E(1)	Dataverse
dvn2015dct-28075	none	0	none	Dataverse
dvn2015dct-28574	none	0	none	Dataverse
dvn2015dct-28583	none	0	none	Dataverse
dvn2015dct-28762	none	0	none	Dataverse
dvn2015dct-28946	none	0	none	Dataverse
dvn2015dct-29244	none	0	none	Dataverse
dvn2015dct-29370	none	0	none	Dataverse
dvn2015dct-Z1A1KE	none	0	none	Dataverse
dvn2015ddi-28075	A(1), O(3), E(1)	5	plus D (1) sameAs E(1)	Dataverse
dvn2015ddi-28574	C(1), O(3), E(1)	5	plus D (1) sameAs E(1)	Dataverse
dvn2015ddi-28583	C(1), O(3), E(1)	5	plus D (1) sameAs E(1)	Dataverse
dvn2015ddi-28762	A(2), O(4), E(1)	7	plus D (1) sameAs E(1)	Dataverse
dvn2015ddi-28946	C(1), O(3), E(1)	5	plus D (1) sameAs E(1)	Dataverse
dvn2015ddi-29244	A(1), C(1), O(3), E(1)	6	plus D (1) sameAs E(1)	Dataverse
dvn2015ddi-29370	O(3), E(1)	4	plus D (1) sameAs E(1)	Dataverse
dvn2015ddi-Z1A1KE	C(1), O(3), E(1)	5	plus D (1) sameAs E(1)	Dataverse
dvn2018dct-0KIRBJ	none	0	none	Dataverse
dvn2018dct-4ICF6W	none	0	none	Dataverse
dvn2018dct-AR4HZI	none	0	none	Dataverse
dvn2018dct-DO3MSH	none	0	none	Dataverse
dvn2018dct-GWTQGU	none	0	none	Dataverse
dvn2018dct-LH3BDN	none	0	none	Dataverse
dvn2018dct-S71F9D	none	0	none	Dataverse
dvn2018dct-SAIK8B	none	0	none	Dataverse
dvn2018dct-SX1B0J	none	0	none	Dataverse
dvn2018dct-VPTJZY	none	0	none	Dataverse
dvn2018dct-YU5JPQ	none	0	none	Dataverse
dvn2018ddi-0KIRBJ	A(1), C(75), O(4), E(1)	81	plus D (1) sameAs E(1)	Dataverse
dvn2018ddi-4ICF6W	A(14), C(6), O(4), E(1)	25	plus D (1) sameAs E(1)	Dataverse
dvn2018ddi-AR4HZI	A(11), C(9), O(4), E(1)	25	plus D (1) sameAs E(1)	Dataverse
dvn2018ddi-DO3MSH	A(15),C(9), O(4), E(1)	29	plus D (1) sameAs E(1)	Dataverse
dvn2018ddi-GWTQGU	C(1), O(3), E(1)	5	plus D (1) sameAs E(1)	Dataverse
dvn2018ddi-LH3BDN	C(1), O(2), E(1)	4	plus D (1) sameAs E(1)	Dataverse
dvn2018ddi-S71F9D	O(2), E(1)	3	plus D (1) sameAs E(1)	Dataverse
dvn2018ddi-SAIK8B	C(1), O(4), E(1)	6	plus D (1) sameAs E(1)	Dataverse
dvn2018ddi-SX1B0J	C(1), O(2), E(1)	4	plus D (1) sameAs E(1)	Dataverse
dvn2018ddi-VPTJZY	A(12), C(10), O(4), E(1	27	plus D (1) sameAs E(1)	Dataverse
dvn2018ddi-YU5JPQ	C(2), O(2), E(1)	5	plus D (1) sameAs E(1)	Dataverse

Figure 1. Harvard's Dataverse sample with validation errors.

Harvard's Dataverse (DVN) proved to hold the most complex metadata standard in the selection as its default format, the *ddi* (Data Description Initiative) comes in two "flavours", the LifeCycle and the codebook. It was only the latter that was tested in validation here, as it was given in the *schemaLocation* of the metadata records. The DVN

search API gives responses only as JSON, with no inherent validation schemas, and there is no OAI-PMH API. So, for ease of comparison and validation, an individual item selection approach was used here to get the metadata records as pure XML with inherent validation schemas, if possible.

This included actively deselecting datasets harvested by DVN from other data providers, to ensure that validation errors encountered do not emanate from other metadata sources than the repository under test. In this case, it leaves us with only 73 items that also have publication year 2018, from which a subset of ten items in descending date order was selected, with representation of different data providers (sources) taking precedence over strict chronology.

For the earlier records from 2014 and 2015 the selection similarly was made from the hit list of a search URL including *metadataSource:"Harvard Dataverse"* as a filter. The selected items were then exported individually into two formats, DDI and Dublin Core (in the extended version dcterms). More specifically, datasets described by *dcterms* and *ddi:codebook2.5* were tested, where *dcterms* is much simpler and lax (e.g. with no mandatory elements, almost no attributes and practically no datatype or content value restraints) in validation than ddi:codebook. As a result, the incidence of validation errors was much higher in the ddi:codebook metadata files. This does not mean that the use of DDI should be discouraged, only that Dataverse as a service provider must work harder to ensure compliance with what is perceived as their default metadata standard.

Another difference between the rendition of metadata for Dataverse datasets in those two formats, is that for *dcterms* metadata files, there is no *schemaLocation* in the files, so to validate them properly, it takes an active effort to find the proper schema. Since *dcterms* has no container (root) element specified, the *dcterms* schema furthermore must be complemented by a new container schema, here *metadataDCT.xsd*, which imports the *dcterms* schema, in order to perform validation. In addition, to validate each Dataverse *dcterms* metadata instance file, one must deselect *dcmi* as the default namespace of element metadata in the file, by adding the prefix *dcmi* to xlmns, e.g. *xmlns:dcmi="http://dublincore.org/documents/dcmi-terms/*.

From Dataverse, 44 metadata records in two metadata formats, *dcterms* (*dct* in the *objectID*) and *ddi codebook 2.5* (Data Description Initiative), of 22 individual single items were sampled and validated. As seen in the table above, all the validation errors pertain to the *ddi* records, the more complex and elaborated of the two metadata standards, as noted above. The most common validation error types for these records are **A**, misplaced xml *attributes*, for example, by the introduction a non-compliant attribute URI in the keyword element, in several instances. **C**-type errors, invalid content, appear with varying frequency, but the extreme occurrence of this error type in one record, objectID *dvn2018ddi0KIRBJ* is mainly due to unescaped parts of html-tags (e.g. /p>) within some descriptive elements, thus not complying with general xml-coding.

No change of metadata standard version was detected between older and more recent metadata records. The *dcterms* records have no inherent *schemaLocation*, so a possible version change would not show explicitly. However they all validate perfectly against the *dcterms.xsd* schema from 2008-02-11. The DDI codebook 2.5 schema[15] is from 2014-01-28, so it has been in use during the whole period covered here.

15 DDI codebook 2.5 schema: http://www.ddialliance.org/Specification/DDI-Codebook/2.5/XMLSchema/codebook.xsd

Figshare

objectID	errorTypes	errorCount	note	repository
Figshare				
fig2018qdc	none	0		Figshare
fig2018datacite	M / O [10]	10		Figshare
fig2018oai_dc	none	0		Figshare
fig2015qdc	none	0		Figshare
fig2015datacite	M / O [10]	10		Figshare
fig2015oai_dc	none	0		Figshare
fig2014qdc	none	0		Figshare
fig2014datacite	M / O [10]	10		Figshare
fig2014oai_dc	none	0		Figshare

Figure 2. Figshare sample with validation errors.

Figshare samples, as already noted, were harvested by means of OAI-PMH feeds in three metadata formats, *qdc*, qualified Dublin Core (a combination of simple DC and two elements from *dcterms, i.e. dcterms:hasPart* and *dcterms:hasVersion*), the simple *oai_dc*, providing also the otherwise absent root element in Dublin Core, and *datacite/ kernel-3*. All in all 90 metadata records of 30 items (with *itemtype* 3 and 4 in the searchURL representing specifically *datasets* and *filesets*).

Both the Dublin Core variants proved to validate without error against the schemas in the OAI-PMH feeds selection, holding the same ten items or records each from the three periods November 2014, June 2015 and November 2018. By contrast, the *oai_datacite* metadata records for the same feeds selection of items has an apparently systematic validation error for the *rights* element, where the parent element (according to the schema) *rightsList* is invariably missing. This might not be a very serious error, though. Paradoxically, it may seem, partly due to its being systematic and uniform, it does not have to affect e.g. the possibility to create a fully functional transformation (XSLT) from the present metadata standard (here DataCite/kernel 3) to a future metadata export format. This could be done by simply "cutting short" the corresponding XPath expression, leaving out the supposed parent element */rightsList* and finding the target child element directly by *//rights*.

For comparison, the metadata records for the same items were later "checked out" individually and exported to *datacite* format, which then proved to be simpler, less inclusive, lacking notably the (optional) *rights* element altogether, but thereby also validating without error against the schema for *DataCite/kernel 3*.

It is particularly noteworthy that there has been no change in metadata standards version used in Figshare during the four years covered here, given the fact that DataCite released its version 4.1 already in October 2017, and v4.0 as early as September 2016, while Figshare is still at *kernel-3*, from July 2013. This illustrates perhaps the generally slow penetration of metadata standards updates in applications, but it does nothing to explain the systematic validation error of the element *rights* in this case, since the parent element *rightsList* was there in *kernel-3* schema and it still is there in the later schemas for v4.0 and v4.1.

Zenodo

objectID	errorTypes	errorCount	note	repository
Zenodo				
zen2018datacite-1745396	none	0	none	Zenodo
zen2018datacite-1745343	none	0	none	Zenodo
zen2018datacite-1745279	none	0	none	Zenodo
zen2018datacite-1745032	none	0	none	Zenodo
zen2018datacite-1745143	none	0	none	Zenodo
zen2018oai_dc-1745143	none	0	none	Zenodo
zen2018datacite-1696038	none	0	none	Zenodo
zen2018datacite-1698621	none	0	none	Zenodo
zen2018oai_dc-1698621	none	0	none	Zenodo
zen2018datacite-1744809	none	0	none	Zenodo
zen2018oai_dc-1744809	none	0	none	Zenodo
zen2018datacite-1742390	none	0	none	Zenodo
zen2018oai_dc-1742390	none	0	none	Zenodo
zen2018datacite-1739702	none	0	none	Zenodo
zen2018oai_dc-1739702	none	0	none	Zenodo
zen2015datacite-15961	none	0	none	Zenodo
zen2015oai_dc-15961	none	0	none	Zenodo
zen2015datacite-16064	none	0	none	Zenodo
zen2015datacite-11611	none	0	none	Zenodo
zen2015oai_dc-11611	none	0	none	Zenodo
zen2015datacite-16191	none	0	none	Zenodo
zen2015oai_dc-16191	none	0	none	Zenodo
zen2015datacite-16282	none	0	none	Zenodo
zen2015oai_dc-16282	none	0	none	Zenodo
zen2014datacite-13234	none	0	none	Zenodo
zen2014oai_dc-13234	none	0	none	Zenodo
zen2014datacite-12942	none	0	none	Zenodo
zen2014oai_dc-12942	none	0	none	Zenodo
zen2014datacite-12646	none	0	none	Zenodo
zen2014oai_dc-12646	none	0	none	Zenodo
zen2014datacite-12727	none	0	none	Zenodo
zen2014oai_dc-12727	none	0	none	Zenodo
zen2014datacite-12873	none	0	none	Zenodo
zen2014oai_dc-12873	none	0	none	Zenodo
zen2014oai_datacite1	C(100), D/E(100)	200	plus C(100) sameAs first C(100)	Zenodo
zen2014oai_dc1	none	0	none	Zenodo
zen2018oai_datacite2	C(7), D/E (7)	14	plus C(7) sameAs first C(7)	Zenodo
zen2018oai_dc2	none	0	none	Zenodo
zen2018oai_datacite3	C(93), D/E(93)	186	plus C(93) sameAs first C(93)	Zenodo
zen2018oai_dc3	none	0	none	Zenodo

Figure 3. Zenodo sample with validation errors.

Zenodo metadata were also harvested by means of OAI-PMH, for two formats, one being oai_datacite, described as the recommended format, containing "the most complete metadata" as the "primary supported export format", which "will always deliver metadata according to the latest available DataCite schema version."[16] The other metadata standard for which records were harvested from Zenodo was simple Dublin Core, *oai_dc*.

Whereas OAI-PMH feeds harvested from Figshare hold only ten items records each, the Zenodo feeds have a full 100 item records. But performing validation on these feeds was somewhat problematic, for several reasons. First, already the fully legitimate URL for harvest, involving a date parameter *from=YYYY-MM-DD*, for some reason creates an extra underscore in the corresponding attribute in the <*request*>-element at the top of the feed, thus: *'from_='*, which must first be removed to make it validate against the general feed schema *OAI-PMH.xsd*. Once the underscore thus has been removed, there is a problem with the validation schema given by: *xsi:schemaLocation="http://schema.datacite.org/oai/oai-1.0/ oai_datacite.xsd"*, which is not to be found, neither in the general OAI-PMH namespace, nor in the indicated *schema.datacite.org* namespace.

A GitHub post from 2011[17] finally indicates the correct reference as: *xsi:schemaLocation="http://schema.datacite.org/oai/oai-1.0/ http://schema.datacite.org/oai/oai-1.0/oai.xsd"*.

After adjusting the references in the metadata records to the proper schema, then, the feeds were validated, with varying results. It should be noted, however, that few if any of the single records in those feeds represent actual datasets; rather the feeds largely contain articles, figures and the like. Whether or not this fact has anything to do with the ensuing validation errors is difficult to say without a more thorough analysis. At this point it may seem, though, that the all but systematic validation errors are more closely related to a failure to reproduce accurately the chosen metadata standard, *datacite/kernel-3*, by the stylesheet involved, */static/xsl/oai2.xsl*, which presently cannot be found for closer analysis. The feed named, *zen2014oai_datacite1* actually had the same two errors in all 100 records.

Since apparently the OAI-PMH feeds from Zenodo offer no parameter for resource type (corresponding to item type in Figshare), a manual search and copying of 35 individual records, specifically for *type=dataset* in two export formats, 20 in *datacite/kernel-4.1*, and 15 in *oai_dc*. These were more or less randomly selected from the years 2014-2015 and 2018. All of these proved to validate perfectly against the inherent schemas, without a single error. This may show, that it is worth the trouble to update your selected metadata standard to the latest on the market, here thus from *datacite/kernel-3* to the more recent *kernel-4.1*? This happened to be the only version change of a metadata standard that was identified in our study, but – apparently – it had little to do with the simple lapse of time. Rather, it seems to be due to the different export functions and production of metadata records in Zenodo, with a less developed OAI-PMH API still using the older version of DataCite, while individual records use an updated version.

16 DataCite schema: http://developers.zenodo.org/#metadata-formats
17 See: https://github.com/datacite/schema/issues/3

Conclusions

The study shows the necessity to store and preserve validation schemas for different metadata standards and versions together with the metadata records and data files in the repositories. The results demonstrate that validation schemas are an important contributor of what the OAIS standard refers to as 'representation information' (CCSDS, 2012).

In some respects the results described for specific repositories are to be expected. Given the fact that the Dublin Core standard and corresponding validation schemas are generally quite lax, with little or no content control, it is hardly surprising that all three repositories manage to produce error free metadata records in *dcterms*, *oai_dc*, and *qdc* formats. More disappointing is the apparent inability to make records comply fully with their own preferred or default metadata standards, DDI or DataCite, and to keep pace with the evolution of metadata standards new versions. Interestingly, Dataverse, which as we saw scored best on the fairmetrics.org test referred to above, in this study seems to have the most problems with validation.

Most of the validation errors found in this study may be more or less easy to fix, with better style sheets for export of metadata. Alternatively, insofar as the errors appear to be systematic, they may simply be bypassed in XPath expressions in the design of transformation files. Nevertheless, it is also problematic that metadata records in some cases found here lack proper, correct *schemaLocations* for their own validation.

Finally, whilst it is not possible to foresee the particular metadata formats that will be used in the future, a useful follow-up study could evaluate the effects of validation errors on transformation efforts more in detail. This would involve developing a simple model transformation (XSLT) of an erroneous original file and the corresponding corrected file (i.e. one that validates against the given schema), to potentially result in an imaginary new metadata standard. Such a model transformation should also be tested on a larger sample of metadata files than those treated in this paper.

In any case, repository managers still have some "running" to do, to catch up with the Red Queen.

References

CCSDS. (2012). Reference model for an Open Archival Information System (OAIS) CCSDS 650.0-M-2 Consultative Committee for Space Data Systems. Retrieved from https://public.ccsds.org/Pubs/650x0m2.pdf

FORCE11. (2016). *The FAIR data principles*. Retrieved from https://www.force11.org/group/fairgroup/fairprinciples

Kingsley, D. (2018). The 'end of the expert': why science needs to be above criticism. In: *Towards cultural change in data management – data stewardship in practice.* (TU Delft 24 May, 2018). doi:10.5281/zenodo.1254830

Li, C. & Sugimoto, S. (2014). *Provenance Description of Metadata using PROV with PREMIS for Long-term Use of Metadata.* Paper presented at DCMI International Conference on Dublin Core and Metadata Applications (Austin TX, 2014). Retrieved from http://dcpapers.dublincore.org/pubs/article/view/3709

Li, C. & Sugimoto, S. (2018). Provenance description of metadata application profiles for long-term maintenance of metadata schemas. *Journal of Documentation, 74*(1), pp.36-61. doi:10.1108/JD-03-2017-0042

Wehrle, D. & Rechert, K. (2018). *Are research data sets FAIR in the long run?* Paper presented at IDCC 2018. Retrieved from http://www.dcc.ac.uk/sites/default/files/documents/IDCC18/PresentationsIDCC18/CWehrle_IDCC2018.pdf